# 《大数据算法综合实践》任务书

**【实验目的】**

本实验旨在通过大规模图数据中三角形计数算法的设计与性能优化，帮助学生深入理解图计算系统的工作原理和性能优化机制，并学会使用图计算框架进行大规模图数据分析和处理。通过此实验，学生将能够掌握图计算的基本概念、编写比较复杂的图算法程序并进行性能调优。

**【实验内容】大规模图数据中三角形计数算法的设计与性能优化**

大数据时代，对关联（图）数据的处理被广泛应用于社交网络、智能交通、移动网络等领域。对图数据的三角形计数被广泛应用于图数据的特征描绘（如聚集系数、联通度等）、社区结构检索、子图匹配、生物网络等应用。

本实验要求在给定服务器平台，以及数据集上实现三角形计数（Triangle Counting，TC）算法，调试并获得最高的性能。三角形的定义是一个包含三个顶点的子图，其中顶点两两相连。例如，以下的无向图中包含 2 个三角形。

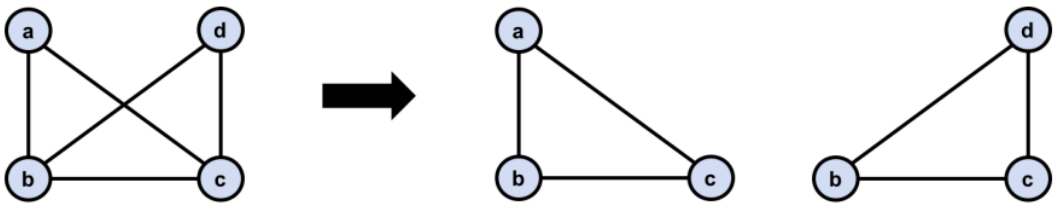


图 1. 三角形计数示例

算法应着重讨论简单无向图的情形，即将重边（multi-edge）看成一条边，同时应不考虑 loop（顶点指向自己的边）。如在上图中，若存在顶点 b 到 c 的两条边，则应忽略其中一条，这样，结果仍然是找到 2 个三角形。在本实验的数据

集中，也应将有向图看成是无向图。在本实验中，同学们需要重点考虑如何在大规模图数据中，精确计算三角形的个数。在真实的数据集中，所考虑的图数据的规模将达到 228~229 顶点和 4-5billion 条边，需要很长的计算时间。因此，大家可以先构造比较简单的数据集用于测试代码功能。

在 Linux 环境下，开发采用 Spark GraphX 或 Pregel 运行时的三角形计数算法。所开发的算法能够充分利用多核资源，以完成给定格式的图数据中三角形的计数。开发环境如下：

- 操作系统：Linux ubuntu 14.04 或 16.04
- 编译器：gcc 或者 g++ 4.8 以上
- Make：GNU make 4.0 以上
- 框架：Spark GraphX、Pregel、或你所熟悉擅长的其它框架

| 数据集 | 说明 |
| --- | --- |
| soc-LiveJournal1 | 顶点数：4.8 million、边数：69 million<br>来源：https://snap.stanford.edu/data/soc-LiveJournal1.html |
| cit-HepPh | 顶点数：34546，边数：421578<br>来源：https://snap.stanford.edu/data/cit-HepPh.html |

注意：

1）以上文件皆为二进制文件，为顺序存放的边表。每条边包含两个顶点（源顶点和目的顶点，各占 4 个字节），边的存储长度为 8 个字节，文件的存储格式为：

源顶点（4 字节，无符号整型）目的顶点（4 字节，无符号整型）源顶点

（4 字节，无符号整型）目的顶点（4 字节，无符号整型）……[EOF]

2）由于 TC 算法要求输入的图为无向图，所以算法应将以上表格中所列出的图均（应视为）无向图处理。

**【实验要求】**

本实验有功能和性能两方面的需求，首先需要能够正确地统计出所给数据集中的所有三角形，其次需要不断优化你所设计的算法（存储优化、通信优化），使得在最终的测试机器上执行时间最短，最终的成绩和你的算法优化效果密切相关。

实验完成后，需要撰写一份详细描述你的算法实现和优化机制的文档，以班级为单位于 6 月 28 日前交东五楼 206 房间。